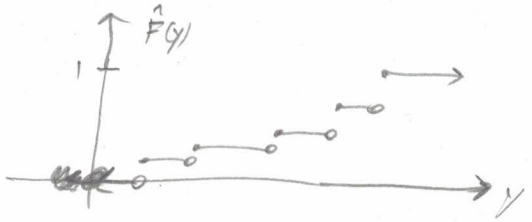
lec 07 MATH 343

This method will work for all parametric models. But what if you don't know the model? (i.e. $Y_i \overset{rv}{\sim}$ model is assumed) — non-parametric

Let's go back to 341. Recall $\hat{F}$, the empirical CDF estimator for $Y_1, \dots, Y_n$ iid.

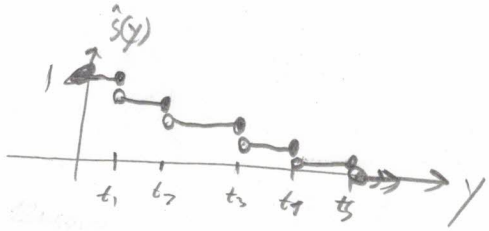
$$\hat{F}(y) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \le y}$$



Thus $\hat{S}$, the empirical survival estimator is just its complement

$$\hat{S}(y) := 1 - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \le y}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{y > Y_i}$$



Let the t_i's be the order statistics of the Y_i's. Assume all unique. Thus, $0 < t_1 < t_2 < \dots < t_n$. Let's tabulate:

time	# still surviving (n_i)	# "at risk" at time of first death event
$t_0 := 0$	$n_1 = n$	$0 = n_{n+1} < n_n < \dots < n_2 < n_1 = n$
t_1	$n_2 = n-1$	
t_2	$n_3 = n-2$	
$\vdots$		
t_n	$n_{n+1} = n-n = 0$	

$\hat{S}(0) = 1$

$$\hat{S}(t_1) = \hat{P}(T > t_1) = \frac{\sum \mathbb{1}_{Y_i > t_1}}{n} = \frac{n_2}{n_1} = \frac{n-1}{n} = 1 - \frac{1}{n}$$

$$\hat{S}(t_2) = \hat{P}(T > t_2) = \hat{P}(T > t_2 \mid T > t_1)\, \hat{P}(T > t_1) = \frac{\sum \mathbb{1}_{Y_i > t_2}}{\sum \mathbb{1}_{Y_i > t_1}} \cdot \frac{n_2}{n_1} = \frac{n_3}{n_2} \cdot \frac{n_2}{n_1} = \frac{n_3}{n_1} = \frac{1}{n} \sum \mathbb{1}_{Y_i > t_2}$$

$$\hat{S}(t_3) = \hat{P}(T > t_3) = \hat{P}(T > t_3 \mid T > t_2)\, \hat{P}(T > t_2 \mid T > t_1)\, \hat{P}(T > t_1) = \frac{n_4}{n_3} \cdot \frac{n_3}{n_2} \cdot \frac{n_2}{n_1} = \frac{n_4}{n_1} = \frac{1}{n} \sum \mathbb{1}_{Y_i > t_3}$$

$$= \frac{n_3 - 1}{n_3} \cdot \frac{n_2 - 1}{n_2} \cdot \frac{n_1 - 1}{n_1} = \left(1 - \frac{1}{n_3}\right)\left(1 - \frac{1}{n_2}\right)\left(1 - \frac{1}{n_1}\right)$$

$$= \left(1 - \frac{1}{n-2}\right)\left(1 - \frac{1}{n-1}\right)\left(1 - \frac{1}{n}\right)$$

$$\hat{S}(t_k) = \prod_{i=k}^{k} \left(1 - \frac{1}{n_i}\right) = \prod_{i=1}^{k} \left(1 - \frac{1}{n-(i-1)}\right)$$

"product limit estimator"

If $j,k \in \{1,\dots,n\}$

If $j<k$, $y \in (t_j, t_k) \Rightarrow \hat{S}(y) = \prod_{\{i: y<t_i\}} \left(1 - \frac{1}{n_i}\right)$

How about inference for a certain value of y e.g. 1 yr?

$\theta := P(Y>1)$, prob. of surviving to 1 yr or more.

$$\hat{\theta} = \hat{S}(1) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i>1} = \prod_{\{i: 1<t_i\}} \left(1 - \frac{1}{n_i}\right)$$

We've seen this before...

Let $B_i := \mathbb{1}_{y_i>1} \Rightarrow B_1,\dots,B_n \overset{iid}{\sim} \text{Bern}(E[\mathbb{1}_{Y_i>1}]) = \text{Bern}(P(Y_i>1)) = \text{Bern}(\theta)$

$$\Rightarrow \hat{\theta} = \bar{B} \overset{\text{by CLT}}{\approx} N\left(E[B_i], \frac{Var(B_i)}{n}\right) = N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

By Slutsky's $\Rightarrow \hat{\theta} \approx N\left(\theta, \frac{\hat{S}(1)(1-\hat{S}(1))}{n}\right)$

With this result, you can plot $\pm$ bands around $\hat{S}(t)$

How about inference for a quantile? Eg. median

$\hat{\theta} = \widehat{Med}(Y) = \underset{y}{\text{argmin}}\left\{\hat{S}(y) \le \frac{1}{2}\right\}$ $\quad Var[\hat{\theta}]$? This is worked out asymptotically but beyond scope of course. Why?

Thm: $\hat{\theta} \approx N(Med(Y), Var[\hat{\theta}])$

For the purposes of this class, use the bootstrap!

Inference for the mean? Use $\bar{Y} \approx N\left(\theta, \frac{S^2}{n}\right)$

Compare 2-samples

$Y_{1,1},\dots,Y_{1,n_1} \overset{iid}{\sim} DGP_1$, $Y_{2,1},\dots,Y_{2,n_2} \overset{iid}{\sim} DGP_2$ Both DGP's Same DGP's $H_0: DGP_1 \neq DGP_2$ $\Rightarrow$ Use 2-Sample K-S test

What if there is non-unique survival times?

No change! However, there is an updated formula for the product limit estimator

Let $n' :=$ # of unique t_i's where $n' \le n$ and $n' < n$ if there is at least one duplicate

time	# of events (d_i)	# still surviving (n_i)
$t_0 = 0$	0	$n_1 = n$
t_1	d_1	$n_2 = n_1 - d_1$
t_2	d_2	$n_3 = n_2 - d_2$
t_3	d_3	$n_4 = n_3 - d_3$
⋮		
$t_{n'}$	$d_{n'}$	$n_{n'+1} = n_{n'} - d_{n'} = 0$

$0 = n_{n'+1} < n_{n'} < \dots < n_2 < n_1 = n$

Note: the last group never enters into the equation

$$\hat{S}(t_3) = \hat{P}(T > t_3 \mid T > t_2)\, \hat{P}(T > t_2 \mid T > t_1)\, \hat{P}(T > t_1)$$

$$= \frac{n_4}{n_3} \cdot \frac{n_3}{n_2} \cdot \frac{n_2}{n_1} = \frac{n_4}{n_1} = \frac{1}{n} \sum \mathbb{1}_{y_i > t_3}$$

$$= \frac{n_3 - d_3}{n_3} \cdot \frac{n_2 - d_2}{n_2} \cdot \frac{n_1 - d_1}{n_1}$$

$$= \left(1 - \frac{d_3}{n_3}\right) \cdot \left(1 - \frac{d_2}{n_2}\right) \cdot \left(1 - \frac{d_1}{n_1}\right)$$

d_i ← # died in this time period

$$\hat{S}(t_K) = \prod_{i=0}^{K-1} \left(1 - \frac{d_i}{n_i}\right)$$

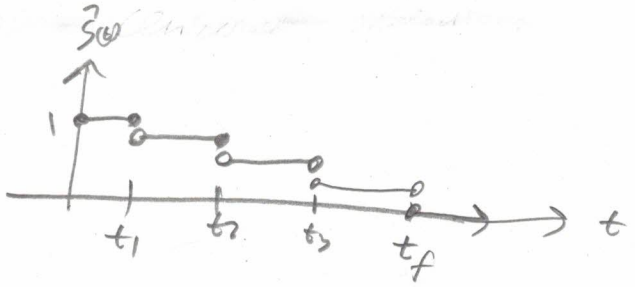
n_i ← # "at risk" or still alive just before t_K

$$\hat{S}(t) = \prod_{\{i : y < t_i\}} \left(1 - \frac{d_i}{n_i}\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{y_i > t}$$

What about censoring?

If we have right-censoring at $t_f := \max(y)$, then what happens?

$$\hat{S}(y) = \frac{1}{n}\mathbb{1}_{y_i > y}$$

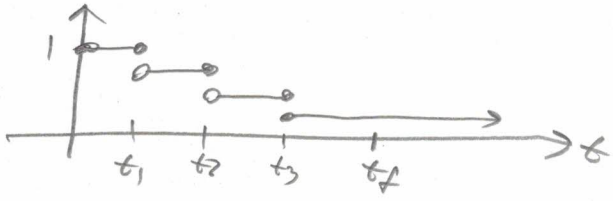


this is clearly wrong!

we know $\hat{S}(y) > 0$ for some $y > t_f$. censored observations are not the same as events!

Solution: just ignore the censoring!

$$\hat{S}(y) = \frac{1}{n}\mathbb{1}_{y_i > 0 \text{ and } c_i = 0}$$



Much more realistic.

Dropping all y_i s.t. $c_i = 1$ from the dataset changes nothing except ...

$\bar{y}$ is now biased downward you deleted lots of real y_i's where $y_i > t_f$ $\forall i$! No good way to get around this. This is why people focus on $\theta = MED[Y]$ and not $E[Y]$ since it's tractable. Others focus on $\theta = \text{restricted mean} := \int_0^{t_f} y f(y)\,dy$ which is a truncated support $[0, t_f] \subset S_y = [0, \infty)$.

$F_Y(t_f)$ but we won't study this